RNA Gene Prediction

Intermediate

Article contents

Stephen R Holbrook, *Lawerence Berkeley National Laboratory, Berkeley, CA, USA*
Richard J Carter, *Lawerence Berkeley National Laboratory, Berkeley, CA, USA*
Richard F Meraz, *Lawerence Berkeley National Laboratory, Berkeley, CA, USA*

The rapid growth of sequenced genomes and discovery of numerous novel RNA species has made the development of methods for the computational identification of genes encoding functional RNA a high priority. Recent developments of algorithms for RNA gene prediction are based on diverse criteria, including location of promoters and terminators, sequence conservation among related genomes, RNA base-pairing and nucleotide composition.

Introduction

0271.1 The promise of genome sequencing has been to reveal the complete set of genes through which the cell performs its functions, as well as regulatory elements that control their expression levels. To a great extent this goal has been reached for protein genes. Powerful computer programs such as GLIMMER, GRAIL and GeneMark have been developed for protein gene discovery and are routinely applied to genomic sequences to identify all open reading frames (ORFs) as potential protein genes. Sequence similarity is used to infer function and confirm gene identity. Hypothetical and unknown proteins are assigned as the products of the remaining genes that lack sequence homology to proteins of known function. The presence of sequence homologs in other organisms confirms the identification of hypothetical proteins even though their function is not assigned. (*See* A0266; A0267.)

0271.2 However, despite the recent rapid expansion in our understanding of the function of ribonucleic acid (RNA) in biological systems, until recently no analogous programs have been developed for the identification of genes encoding functional RNAs (fRNAs) in genomic sequences. This omission was due to both the conceptual underappreciation of the number and significance of RNA genes in genomic sequences and technical difficulties due to lack of signals within and around RNA genes in comparison to protein genes. (*See* A0008; A0985.)

0271.3 **Figure 1** shows these differences for procaryotic genes. As shown, protein genes incorporate several signals that can be used in their recognition. These include the consensus ribosome binding (Shine–Dalgarno) site, the appropriately spaced start and stop codons, and a triplet-encoded ORF based on the genetic code and corresponding to the organism's codon usage frequencies. On the other hand, RNA genes contain none of these signals.

0271.4 Computational prediction of fRNAs in genomic sequences would allow experimental testing of expression levels, functional assay by deletion or mutagenesis, and structural analysis. These untranslated fRNAs have also been referred to as noncoding (ncRNA), small RNAs (sRNA, smRNA), untranslated and small nonmessenger (snmRNA). Here they are referred to as fRNAs.

0271.5 Two recent experimental studies have identified an unexpectedly large and diverse population of expressed and presumably fRNA molecules. First, an expressed sequence tag (EST)-based experimental technique was used to generate complementary deoxyribonucleic acid (cDNA) libraries from the small fraction (50–500 nucleotides) of total RNA isolated from mouse brain (Huttenhofer *et al.*, 2001). This approach resulted in the identification of a total of 201 different expressed RNA species potentially encoding fRNA species. Of these RNAs, 113 were identified as small nucleolar RNAs (snoRNAs), guiding modification of ribosomal, snRNA or other RNA modifications. Most of the remainder consisted of novel RNAs of unknown function. (*See* A0357.)

0271.6 Several groups have used a range of biochemical techniques, including cDNA cloning and cloning from

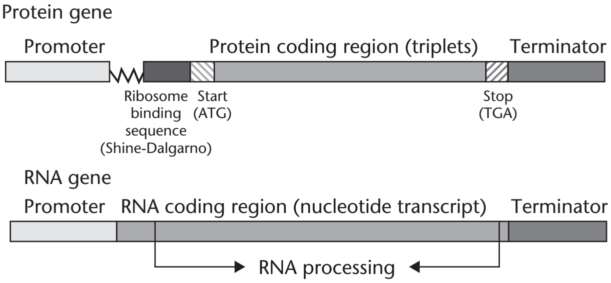

Protein gene
Promoter
Protein coding region (triplets)
Terminator
Ribosome binding sequence (Shine-Dalgarno)
Start (ATG)
Stop (TGA)
RNA gene
Promoter
RNA coding region (nucleotide transcript)
Terminator
RNA processing

Figure 1 Comparison of RNA and protein genes in procaryotes. In the RNA gene, there are no ribosome binding sites, no start or stop codons, and no triplet code. 0271.F001

selected fractions of total RNA to discover the presence of micro-RNAs (miRNAs; for a review see Ruvkin (2001)). These miRNAs act in a variety of roles, including development and regulation, and were identified in *Caenorhabditis elegans*, *Drosophila melanogaster* and HeLa cells. It is expected that these miRNAs modulate the translation or stability of messenger RNAs (mRNAs; Lau *et al.*, 2001). (*See* A0344; A0349.)

Au. Qu. from editor: Pl. chk if change okay.

Computational Approaches

Identification of well-characterized RNA species

0271.7 Many fRNAs are shared among organisms, such as transfer RNA (tRNA) and ribosomal RNA (rRNA), ribonuclease P (RNase P) RNA, snoRNAs (eucaryotes and archaea) and tmRNAs (bacteria). These known RNAs can usually be identified in genomes by sequence and/or structure similarity and conservation. Sophisticated programs such as tRNAscan-SE and Snoscan utilize conserved sequence and/or structure patterns, covariance models and stochastic context-free grammars (SCFGs; Eddy and Durbin, 1994) to accurately and automatically find tRNAs (Lowe and Eddy, 1997) and snoRNAs (Lowe and Eddy, 1999; Omer *et al.*, 2000) in genomes. These methods were developed to find new members of well-characterized RNA types, but are not applicable to the identification of novel or poorly characterized RNA species. (*See* A0274.)

0271.8 Formal Bayesian probabilistic models have been introduced as tools to identify complicated consensus features in biological sequences. Hidden Markov models (HMM) are probably the best known of these approaches. Another class of model, the covariance model, is able to capture both primary consensus and secondary structure information through the use of SCFGs. Much like sequence profiles, covariance models are constructed from multiple sequence alignments. (*See* A0851.)

0271.9 In the tRNAscan-SE program (Lowe and Eddy, 1997), sequences are searched against a given covariance model using a three-dimensional dynamic programming algorithm, similar to a Smith–Waterman alignment but including base-pairing terms also. RNA covariance models have the advantages of high sensitivity, high specificity, and general applicability to any RNA sequence family of interest. Using these general tools, the search for a tRNA takes three steps. Firstly, the DNA is screened for the presence of a short, intergenic promoter sequence that is found in the T and D arms of tRNA. This is followed by a search for stem-loop structures in the location of the promoter. The second step involves calculating a log-odds score for conserved sequences and the distance between. The final stage parses the output from these programs and undertakes a probabilistic search for tRNA. A secondary structure prediction of any putative matches will reveal the presence of an anticodon region. (*See* A0254; A0263.)

Snoscan (Lowe and Eddy, 1999) is a program that recognizes methylation guide snoRNAs in archaeal genomes. The program identifies six components characteristic of the class of fRNA: box D; box C; a region of sequence complementary to rRNA; box D9 if the rRNA complementary region is not directly adjacent to box D; the predicted methylation site within the rRNA based on the complementary region; and the terminal stem base pairings, if present. The program also takes into account the relative distance between identified features within the snoRNA, information that is useful in reducing the rate of false positives. 0271.10

Identification of novel fRNAs

Recently, a number of studies have been undertaken to find novel RNA genes using gene boundary prediction (Olivas *et al.*, 1997; Argaman *et al.*, 2001), comparative genomics (Argaman *et al.*, 2001; Wassarman *et al.*, 2001), a combination of comparative sequence analysis and probabilistic models of nucleotide mutation bias in regions of conserved secondary structure (Rivas *et al.*, 2001), and contrasting sequence and structural patterns between known RNA genes and noncoding sequences within genomes (Carter *et al.*, 2001). The various approaches to computational identification of novel RNA genes are summarized in **Table 1** and described in the following sections. 0271.11

Gene boundary prediction

The prediction of potential RNA promoters and terminators has been used to narrow the search for potential RNAs. Parker and coworkers used a genomics guided-search technique to find novel RNA genes (Olivas *et al.*, 1997). Two strategies were used in this study. First, strong RNA polymerase III sites were identified by sequence, and transcripts from these sites were probed experimentally. Second, large gaps between predicted ORFs were analyzed for RNA expression. The first method identified a new, but nonessential, 170-nucleotide noncoding RNA, and the second method found 15 RNA transcripts, one of which appeared to be an snoRNA. While this approach was laborious and not comprehensive, it did show the presence of previously unidentified RNAs in the yeast genome. Argaman *et al.* (2001) used promoter and terminator prediction methods combined with comparative genomics to predict potential RNA coding regions in *Escherichia coli*. (*See* A0269.) 0271.12

0271.T001 **Table 1** Computational approaches for the identification of novel RNA genes

Basis	Method	Organism	Number predicted	Reference
Boundary prediction (promoter–terminator patterns)	CSA	Yeast *E. coli*	16 24	Olivas *et al.* (1997) Argaman *et al.* (2001)
Comparative genomics – sequence conservation	CSA	*E. coli*	23 24	Wassarman *et al.* (2001) Argaman *et al.* (2001)
Stability of RNA (GC content)	NCS	Archaea	NA >200	Schattner (2001) Omer *et al.* (2000)
Nucleotide mutation bias in conserved secondary structure	SCFG/HMM	*E. coli*	275	Rivas *et al.* (2001)
Comparison of known RNAs to noncoding non-conserved intergenic sequences	NN SVM	Archaea *E. coli*	370	Carter *et al.* (2001)

NN: neural networks; SVM: support vector machines; SCFG: stochastic context-free grammars; HMM: hidden Markov models; CSA: comparative sequence analysis (BlastN); NCS: nucleotide composition statistics.

0271.13 Currently, the polymerase binding sites of RNA promoters and terminator sites are not unambiguously predicted in bacteria, archaea or eucaryotes. In addition, due to the variable distance between promoters and start sites, RNA processing, and the presence of introns in eucaryotes, these signals alone do not define the fRNA very well. Progress in this area, combined with other approaches, will, however, be important in improving predictive methods. For example, an improved algorithm for prediction of rho-independent terminators in procaryotes has been reported recently (Lesnik *et al.*, 2001).

Comparative genomics

0271.14 Wassarman *et al.* (2001) have used a comparative genomics study to identify 19 novel RNA genes in *E. coli*. They examined intergenic sequences of more than 180 nucleotides (nt) and then carried out a BLAST search against a set of bacterial genomes. Those that had a high degree of conservation over at least 80 nt were further examined. After screening to remove those that were possibly ORFs or promoters, the remaining sequences were tested using microarray data and traditional biochemical methods. The authors speculate that the comparative genomics approach may be easily applied to other organisms, although it is noted that a high degree of conservation does not necessarily infer the presence of an RNA gene, but may instead be a protein binding site, control element, insertion sequence or even a small ORF. (*See* A0253.)

0271.15 Another comparative genomics study of *E. coli* (Argaman *et al.*, 2001) has examined intergenic sequences and additionally attempted to identify sites of transcription initiation or termination within them. If the distance between the initiation and termination sites was between 50 and 500 nucleotides then a BLAST search of the sequence was undertaken using three bacterial genomes for comparison. This produced 24 potential RNA genes, of which 14 were biochemically verified. It is hypothesized that, provided the genomic and sequence features characteristic of an fRNA can be defined in an explicit manner, then an algorithmic approach can be used to find fRNA in higher organisms.

An analysis of the *E. coli* genome has been 0271.16
undertaken using a program called QRNA (Rivas *et al.*, 2001) that combines comparative genomics and probabilistic methods. The program requires as input a set of sequence alignments from closely related organisms and then classifies the sequence as an fRNA, a protein coding sequence or an uncorrelated sequence region. The hypothesis is that conserved sequence regions will show a pattern of mutation more consistent with a probabilistic model of covariation of nucleotides within base-paired secondary structure rather than an analogous model of conservation of an encoded amino acid sequence (protein coding) or the null model of uncorrelated, position-independent mutation. While this method is entirely general, requiring no prior knowledge about the fRNAs of the respective genomes, it is limited to detecting fRNAs with conserved intramolecular structure. The method was used to identify 275 putative fRNAs in *E. coli*, a number of which have been verified with biochemical methods (Rivas *et al.*, 2001). Testing indicated a true-positive prediction accuracy of approximately 85% for classification of RNAs in *E. coli*.

Compositional and structural characterization

Defining the features characteristic of fRNAs has led 0271.17
to a great deal of recent research. A computational approach to the identification of RNA genes has been applied to archaeal genomes (Omer *et al.*, 2000; Schattner, 2001) based on the increased G+C content

of RNA genes in hyperthermophilic organisms such as *Methanococcus jannaschii* that have an AT-rich overall genome composition. While this method appears to have predictive value, its use is restricted to organisms with AT-rich genome compositions. (*See* A0260.)

0271.18 Carter *et al.* used machine-learning methods to predict the presence of novel RNA genes in the *E. coli* genome and several other bacterial and archaeal genomes. In order to discriminate fRNA from background genomic sequence, a number of neural networks were trained. The first neural network was trained on the single and dinucleotide composition of known fRNAs genes and a data set of presumed nonfunctional genomic sequence. A second network was trained using 'structural motifs'; these included the well-known sequence motifs UNCG, GNRA and CUYG found in RNA tetraloops, the AAR subsequence of the tetraloop receptor motif, and the DNA sequence CTAG, which occurs rarely in bacterial protein genes and noncoding regions compared with RNA genes. The final parameter of this set was the calculated free energy of folding. When a third voting-network was trained on the combined results of the previous networks, it achieved an overall predictive accuracy of over 90% in bacterial genomes, over 95% in hyperthermophilic archaeal genomes, and was successfully able to predict a number of recently identified fRNAs that were not included in the training sets. (*See* A0268.)

Conclusion

0271.19 It is evident from the work of a number of laboratories as of the year 2002 that RNA gene prediction is an area of intense research and that the techniques are improving substantially. It is hoped that RNA gene prediction can become as reliable and have the same level of confidence as genomic protein prediction methods, and that new biological pathways, incorporating novel regulatory, catalytic or structural RNAs, will be identified.

0271.20 Much of this recent work has concentrated on bacterial genomes, especially *E. coli*. This is mainly because of the wealth of available information about that genome, both in terms of sequence and biological function. A number of groups have estimated that there are probably a further 300 RNA genes to be found in *E. coli*. These same techniques are now being applied to higher organisms, especially the human genome, which may contain thousands of novel RNA genes waiting to be discovered.

References

Argaman L, Hershberg R, Vogel J, *et al.* (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli. Current Biology: CB* **11**: 941–950.

Carter RJ, Dubchak I and Holbrook SR (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research* **29**(19): 3928–3938.

Eddy SR and Durbin R (1994) RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**: 2079–2088.

Huttenhofer A, Kiefmann M, Meier-Ewert S, *et al.* (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO Journal* **20**: 2943–2953.

Lau NC, Lim LP, Weinstein EG and Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science* **294**: 858–862.

Lesnik EA, Sampath R, Levene HB, *et al.* (2001) Prediction of rho-independent transcriptional terminators in *Escherichia coli. Nucleic Acids Research* **29**: 3583–3594.

Lowe T and Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**: 955–964.

Lowe TM and Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. *Science* **283**: 1168–1171.

Olivas WM, Muhlrad D and Parker R (1997) Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Research* **25**: 4619–4625.

Omer AD, Lowe TM, Russell AG, *et al.* (2000) Homologues of snoRNAs in Archaea. *Science* **288**: 517–522.

Rivas E, Klein RJ, Jones TA and Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current Biology: CB* **11**: 1369–1373.

Ruvkin G (2001) Glimpses of a tiny RNA world. *Science* **294**: 797–799.

Schattner P (2001) Searching for RNA genes using base-composition statistics. *International Conference on Intelligent Systems for Molecular Biology; ISMB 2001, Copenhagen, Denmark.*

Wassarman KM, Repoila F, Rosenow C, Storz G and Gottesman S (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes and Development* **15**: 1637–1651.

Further Reading

Szymanski M and Barciszewski J (2002) Beyond the proteome: non-coding regulatory RNAs. *Genome Biology* 1–8.

Eddy SR (2002) Computational genomics of noncoding RNA genes. *Cell* **109**: 137–140.

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**: 919–929.

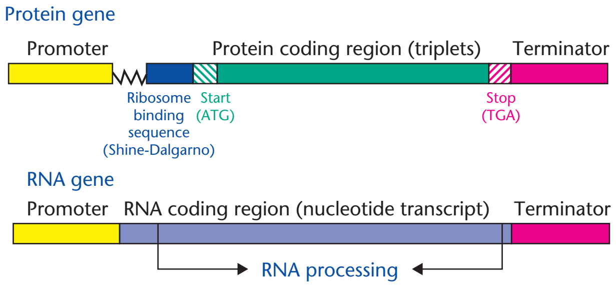
Storz G (2002) An expanding universe of noncoding RNAs. *Science* **296**: 1260–1263.

Regalia M, Rosenblad MA and Samuelsson T (2002) Prediction of signal recognition particle RNA genes. *Nucleic Acids Research* **30**: 3368–3377.

Laslett D, Canback B and Andersson S (2002) BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Research* **30**: 3449–3453.

Schattner P (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Research* **30**: 2076–2082.

Chen S *et al.* (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *BioSystems* **65**: 157–177.


Protein gene
Promoter
Protein coding region (triplets)
Terminator
Ribosome binding sequence (Shine-Dalgarno)
Start (ATG)
Stop (TGA)
RNA gene
Promoter
RNA coding region (nucleotide transcript)
Terminator
RNA processing

0271.F001 **Figure 1** Comparison of RNA and protein genes in prokaryotes. In the RNA gene, there are no ribosome binding sites, no start or stop codons, and no triplet code.

Keywords

functional RNA, gene finding, genomics, bioinformatics, machine learning